HMP WGS Read Mapping

[1]The Genome Institute, Washington University School of Medicine

[2]Broad Institute of MIT and Harvard

Authors: John Martin[1], Sarah Young[2], Makedonka Mitreva[1]

Version: 1.0c

Effective Date:

1 Abstract

2 Introduction

This SOP describes the procedure for mapping reads to reference genomes for the HMP WGS data.

3 Requirements

4 Procedure

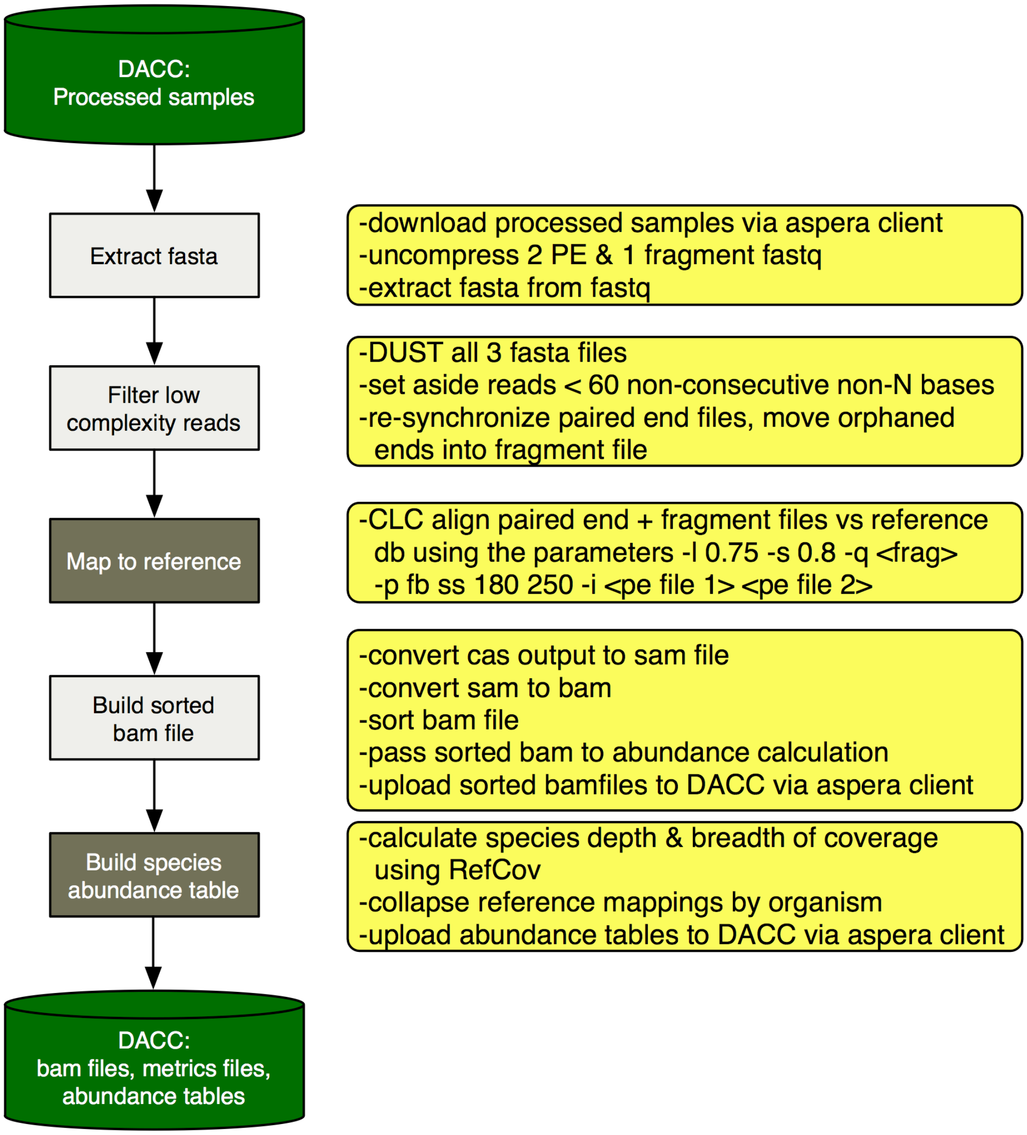
Figure 1 shows an overview of the entire read mapping process.

4.1 Extract fasta

- The process of read mapping to reference genomes for HMP WGS data began with the download of processed reads from **http://hmpdacc.org/HMASM/**. Each sample download was comprised of 2 mated pair fastq files, and 1 fragment fastq file.

4.2 Filter low complexity reads

- Fasta format sequences were extracted from each file, and subjected to a low-complexity screen using the ‘DUST’ program (distributed with NCBI blast). DUST masks out all bases that have low compositional complexity, and "can eliminate statistically significant but biologically uninteresting reports from the output" (For more information, see http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml#filter).
- We considered any read with fewer than 60 non-masked bases (not necessarily consecutive) to be of low complexity, and thus discarded it from the final set. In cases where one end of a paired end set of mates was found to be of low complexity, and the other end was not, the orphaned (but good quality) read was removed from the paired end file, and moved into the fragment read file.


DACC:
Processed samples
Extract fasta
Filter low complexity reads
Map to reference
Build sorted bam file
Build species abundance table
DACC:
bam files, metrics files, abundance tables

-download processed samples via aspera client
-uncompress 2 PE & 1 fragment fastq
-extract fasta from fastq

-DUST all 3 fasta files
-set aside reads < 60 non-consecutive non-N bases
-re-synchronize paired end files, move orphaned ends into fragment file

-CLC align paired end + fragment files vs reference db using the parameters -l 0.75 -s 0.8 -q <frag> -p fb ss 180 250 -i <pe file 1> <pe file 2>

-convert cas output to sam file
-convert sam to bam
-sort bam file
-pass sorted bam to abundance calculation
-upload sorted bamfiles to DACC via aspera client

-calculate species depth & breadth of coverage using RefCov
-collapse reference mappings by organism
-upload abundance tables to DACC via aspera client

Figure 1. Overview of the read mapping process

4.3 Map to reference

- The processed reads not marked as low-complexity were then aligned to our reference genomes database, available at **http://hmpdacc.org/HMREFG/**, using the aligner clc_ref_assemble_long (part of the CLC Assembly Cell package available from CLC bio @ http://www.clcbio.com/) with the parameters "-lengthfraction 0.75 -similarity 0.8 -p fb ss 180 250"

- note: -p fb ss 180 250 sets paired end information, 'fb' indicates that the first read is in the 'f'orward orientation, and the second is in the 'b'ackward orientation (i.e. facing each other)
- the 'ss 180 250' part informs the program to expect the 's'tart to 's'tart (i.e. the far ends, since they are facing each other) distance between the reads to be from 180-250bp in length).
- Both the paired end set and the fragment file were aligned in a single execution of the software.

4.4 Build sorted BAM file

- The immediate results from the aligner were in CAS format, and were converted into BAM format using the script 'castosam' for downstream analysis (the 'castosam' program is included in the CLC bio aligner ("CLC Assembly Cell") distribution).

4.5 Build species abundance table

- After mapping to the reference genomes database, minimal metrics were collected using samtools flagstats. We harvested the total number of reads passing our low-complexity filter and the total number of reads mapping to the database. In total we attempted to map 38,415,726,108 high-quality microbial reads, known to be free of low-complexity sequence, spread over 754 samples, against our reference database. We found 22,113,712,465 hits (~57.6%) using our alignment cutoff of 80% identity over 75% of the length of the query.

- Additionally we built strain abundance tables on a per-sample basis, showing the depth and breadth of coverage of each strain in our database. We reported all strains meeting a coverage cutoff of 0.01x depth across 1% breadth of each strain's reference genome.

- Initial coverage statistics were generated using the Ref-Cov program (internal, WUGI software by Todd Wylie & Jason Walker), and the coverage per each sequence record was collapsed such that we had a cumulative depth and breadth of coverage for each strain in the reference database (many draft strains are comprised of multiple contigs that needed to be collapsed & considered as a single organismal entity for the final report).

5 Implementation

The following files, along with md5 fingerprints for download validation, are available for each sample on the DACC website, at http://hmpdacc.org/HMSCP/

- mapping bam files
- minimal metrics files
- sample abundance tables

6 Discussion

7 Related Documents & References

8 Revision History

Version	Author/Reviewer	Date	Change Made
1.0	John Martin, Sarah Young, and Makedonka Mitreva		Establish SOP
1.0c		09/20/2011	Converted to standard template